Variance Estimation in US Census Data from 1960-2000

Kathryn M. Coursolle, Lara L. Cleveland, & Steven Ruggles

MPC Minnesota Population Center
University of Minnesota
Driven to Discover℠

Introduction

- IPUMS microdata are among the most heavily used data sources for social, historical, demographic, and policy research.
- Census microdata feature complex sample designs that are clustered within households and incorporate stratification.
- Yet, researchers often calculate standard errors utilizing methods designed for simple random samples.
- Failure to adjust for clustering and stratification in the sample design may lead to incorrect standard errors and invalid statistical inferences (Davern & Strief; Kish, 1995; Lohr, 2000).

Research Objectives

- Using decennial census data from 1960-2000 and American Community Survey (ACS) data from 2004-2010 we evaluate the impact of sample design on standard error estimates.
- We compare standard error estimates under the assumption of simple random sampling to variance estimates accounting for clustering and strata using Taylor series linearization.
- In the 2005-2010 ACS samples we also compare standard error estimates to the Census Bureau's subsample replicate weights.

Data

- Decennial census data from 1960-2000 and American Community Survey data from 2004-2010 from IPUMS.org

Stratification in IPUMS-USA Samples from 1960-2000

Census Year	Stratification Characteristics: U.S. Census Public Use Microdata Files	# of strata IPUMS	IPUMS USA Strata Characteristics
1960	Within small area: household size, home ownership, race, group quarters residence (used procedure to combine cells that represented fewer than 50 people in the population)	38	Uses same characteristics for classification (without geography)
1970	Within geographic area: home ownership, race, sex of head, household size, presence of own children, inmate status, other group quarters residence	75	Uses same characteristics for classification (without geography)
1980	By state: 102 strata based on race, Spanish origin, home ownership, presence of own children, group quarters, migration status	51	Similar classification based on race, Spanish origin, home ownership, presence of own children (without geography)
1990	By geographic area: 312 strata for occupied households presence of own children, race, Spanish origin, home ownership, grouped age; separate stratification characteristics for vacant houses (9 strata) and group quarters (56 strata)	119	Similar classification based on presence of children, race (some collapsed categories), Spanish origin, home ownership (without geography)
2000	By geographic area: thousands of strata for occupied households based on presence of own children, detailed race, Spanish origin, home ownership, age of oldest household member; separate configurations for vacant housing units (12) and group quarters (2840)	131	Similar but condensed configuration based on presence of children, collapsed race categories, Spanish origin, home ownership (without geography)

Stratification in the ACS

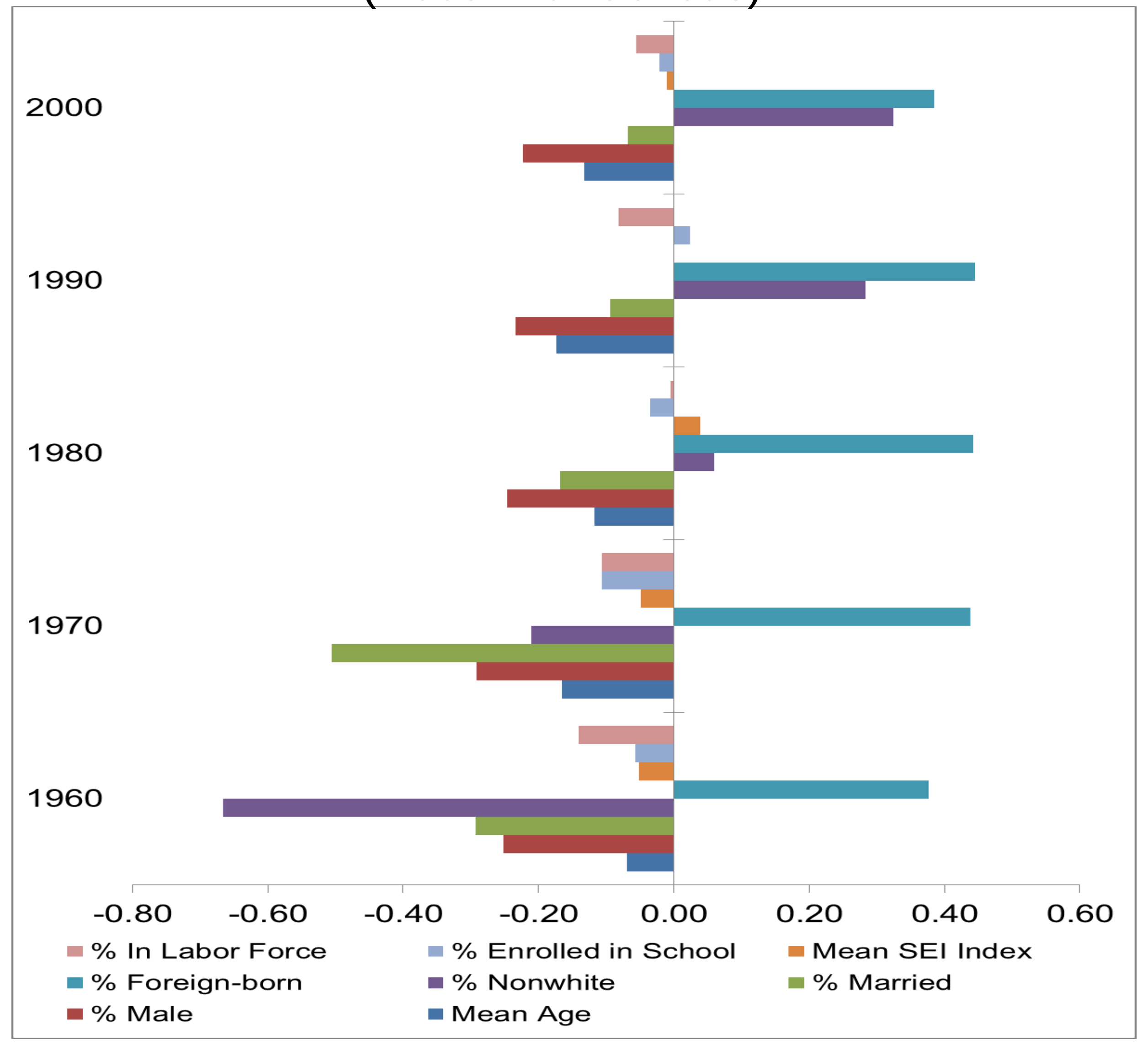
- The ACS is administered to a systematic sample of households by county.
- IPUMS-USA strata use PUMA geographic areas as a proxy for implicit geographic stratification in the data.

Subsample Replicate Weights in the ACS

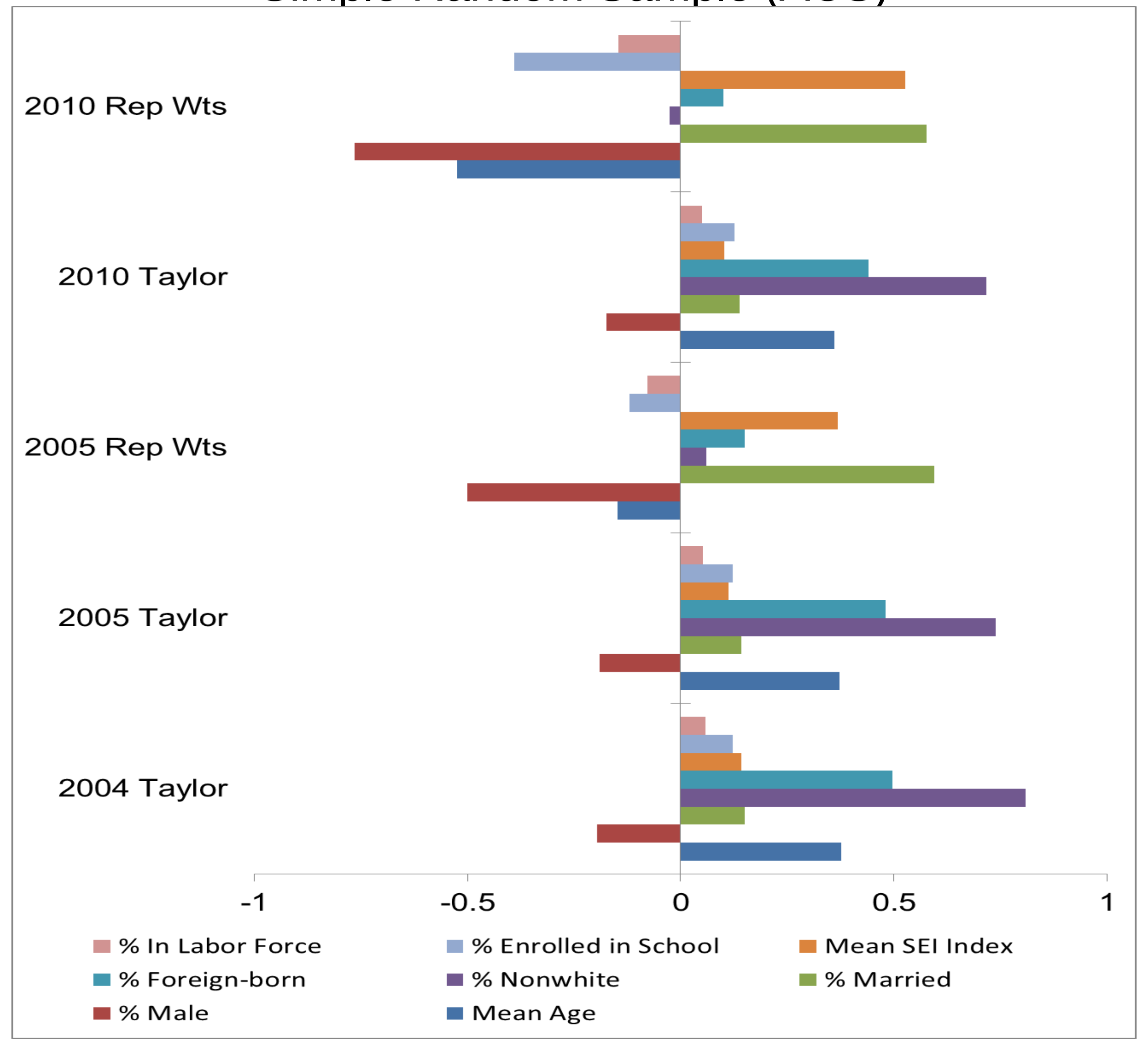
- 80 separate replicate weights produced by the Census Bureau at the household and person levels were added to the ACS in 2005.
- These weights allow the sample to mimic multiple samples, which can produce more informed standard error estimates and reflect relevant sample design information.

Results

Difference in Ratio of Standard Error Adjusting for Clustering and Strata Relative to Simple Random Sample (Decennial Census)



Difference in Ratio of Standard Error Adjusting for Clustering and Pseudo Strata and Using Replicate Weights Relative to Simple Random Sample (ACS)



Conclusion

USA Decennial Samples:
Taylor series standard error estimates are smaller than SRS estimates for most variables but larger for variables that tend to be correlated within households, especially variables not included in strata design such as foreign-born. Use of strata and cluster increases precision by incorporating sample design characteristics. Use of cluster is essential for some variables.

ACS Samples
Taylor series estimates with cluster and geographic pseudo-strata are often larger than standard errors assuming a simple random sample. These estimates are similar to, if generally more conservative than, results using replicate weights and require less computing time.

References

- Davern, M. & Strief, J. "IPUMS User Note: Issues Concerning the Calculation of Standard Errors (i.e., variance estimation) Using IPUMS Data Products" Ipums.org: http://usa.ipums.org/usa/resources/complex_survey_vars/UserNote_Variance.pdf
- Kish, L. 1995. *Survey Sampling*. Wiley Classics Library Edition. New York, NY: Wiley and Sons.
- Lohr, S. 2000. *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Acknowledgements

The authors would like to thank Michel Boudreaux for his helpful suggestions. Research for this project was supported by the National Institutes of Health (grant number R01HD043392). Address correspondence to Kathryn M. Coursolle, Minnesota Population Center (MPC), University of Minnesota, 50 Willey Hall, 225 19th Avenue South, Minneapolis, MN 55455, e-mail: kathrync@umn.edu.